Abstract

This thesis presents a methodology for understanding the performance and scalability of clustering algorithms on parallel computers and hence it explains the scalability analysis of clustering algorithms. We here demonstrate the efficiency of proposed algorithms and show how it can guide the development of better parallel algorithms. We present a new highly scalable parallel architecture for large data that has been utilizing the intrinsic capabilities of a multi-core processor to a limited extent.

The current multi-core architectures have become popular due to performance and efficient processing of multiple tasks simultaneously. The design of parallel algorithm and performance measurement is the major issue on multi-core environment. If one wishes to execute a single application faster, then the application must be divided into subtasks or threads to deliver the desired result. Parallel computing is a critical component of the computing technology of the 90s, and it is likely to have as much impact over the next twenty years as microprocessors have had over the past twenty years. Multi-core technology means having more than one core inside a single chip. This opens a way to parallel computation, where multiple parts of a program are executed in parallel at same time. The factor motivating the design of parallel algorithm for a multi-core system is the performance. An important aspect of performance analysis is the study of how algorithm performance varies with parameters such as problem size, number of processors, and speedup, in particular to evaluate the scalability of a parallel algorithm, that is, how effectively it can use an increased number of processors. The software development tools must abstract these variations so that software performance continues to obtain the benefits of the Moore's law.

Data mining technology is the only best method used in extracting meaningful information from large and unorganized databanks. Due to advances in information technology and high performance computing, very large data sets are becoming available in many scientific disciplines. The rate of production of such data far outstrips our ability to analyze them manually. For example, a computational simulation can generate tera-bytes of data within a few hours, whereas human analysts may take several weeks to analyze these data sets. As a result, there is an increasing interest in various scientific communities to explore the use of emerging data mining techniques for the analysis of these large data sets. Drastic advancement in scientific computing and worldwide communication over wired and wireless networks have resulted in the

generation of enormous amount of data. This will lead to develop parallel data mining technology. Parallel data mining technology involves parallel architecture and parallel algorithm. Data mining techniques have been applied to scientific datasets, with fields such as remote sensing, astronomy, biology, physics and chemistry, providing a rich environment for the practice of these techniques. Efficient parallel clustering algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in scientific data analyses.

Clustering has been used extensively for more than forty years in data mining field and across many disciplines due to its broad applications. The clustering algorithms can be categorized into fuzzy clustering, partitional clustering, hierarchical clustering, density-based clustering, evolutionary clustering and so on. In these methods, hierarchical and partitional clustering algorithms are two primary approaches of increasing interest in research communities. Recently, large data clustering has been extensively studied in many areas, including statistics, machine learning, pattern recognition, and image processing. In the areas, the scalability of clustering methods and the techniques for big data clustering, much active research has been done. To overcome the problems that occurred in large data clustering different methods have been introduced, including initialization by clustering a sample of the data and using an initial crude partitioning of the entire data set. However, the most prominent representatives are partitioning clustering methods such as CLARANS, hierarchical clustering methods such as BIRCH, grid clustering methods such as STING and WAVECLUSTER. Firefly clustering algorithm is one of the recent swarm intelligence methods developed by Yang in 2008 and is a kind of stochastic, nature-inspired, meta-heuristic algorithm that can be applied for solving the hardest optimization problems (also NP-hard problems). Genetic algorithms are randomized search and optimization techniques guided by the principles of evolution and natural genetics, having a large amount of implicit parallelism.

Every method has got its advantages and shortcomings. They may not be suitable for processing very large data. Sometimes it is very difficult to acquire both accuracy and efficiency in a clustering algorithm of a large dataset. The two parameters never complement each other. In order to overcome this drawback and hence to process massive data sets the power of a single computer is not enough. Parallel clustering is the key technique which is highly scalable and in low cost to do clustering in efficient way**.** However, the challenge continues in dealing with large data, because most of the algorithms are compatible only with small data. However, the existing clustering algorithms either handle different data types with inefficiency in handling large data or handle large data with limitations in considering numeric attributes. Hence, parallel clustering

has come into picture to provide crucial contribution towards clustering large data. This insists the need of having scalable parallel clustering to solve the aforesaid problems.

Scalability of discovered knowledge processing is widely used for pure parallel computer systems, and focuses on improving the performance when increasing the number of computing processors. The primary objective of scalability analysis is to determine how well a system can effectively work on larger issues as per the increase in its size. In general, scalability measures the capability of a system to maintain or increase the performance with the increase of the problem or the computer size increases. The best solution which adopts in the same application with the same algorithm may have different results in scalabilities. It is a possibility that the implementation of an algorithm in one system is scalable but the same implementation on another system with different architecture does not give much scalability.

In this research work, we can apply algorithms in different multi cores and also tests in varying number of data sizes. Increased availability and the extensive demand of multi-core processors force us to re-design the existing algorithms and applications so as to exploit the available computational power from multiple cores. For the effective utilization of the intrinsic capabilities of a multi-core processor, the software application must be able to execute tasks in parallel using all the available CPUs. In order to achieve this we can use the mechanism of Fork/Join model in java programming Since the Java threads are run by OS threads, multiple threads run on different cores by default. This is the most effective architectural method to achieve good parallel performance and efficiency. Fork/Join object in Java is able to handle the divide and conquer algorithms in efficiency attainable practice. Also Fork/Join is scalable in the sense that it will run on as many cores as they are supported by the processor chip, dependent on the amount of parallelism and the number of cores significant speedup with this development. One important point to be noted in this framework is that ideally no worker thread is idle in this methodology. They implement a work-stealing algorithm in which the idle workers steal the work from those workers who are busy. The number of worker threads in a Fork/Join pool is generally upper-bounded by the number of cores in the system. The Fork/Join Framework enhances multithreaded programming in two important ways. First, it simplifies the creation and the use of multiple threads. Second, it automatically makes use of multiple processors.

In the present work, the main aim is to make a clustering algorithm that utilizes maximum capabilities of a regular multi-core PC to cluster the dataset as fast as possible while resulting in acceptable quality of clusters. In order to overcome the drawbacks of the existing techniques, two novel scalable algorithms have been designed and implemented, namely;

- ✓ **Scalable Parallel Clustering Approach for Large Data using Genetic Possibilistic Fuzzy C-Means Algorithm**

The parallel architecture is developed by including the techniques like, splitting the input data, clustering each subset of data and merging to optimal final clustering. The initial stage of the proposed method is dividing the input large dataset into n number of dataset according to the n number of cores available in the system, then clustering of subsets of data, PFCM clustering, which is a hybridization of Possibilistic C-Means (PCM) and Fuzzy C-Means (FCM) that often avoids various problems of PCM, FCM and FPCM used. For the optimal merging process, Firefly-based clustering will be modified and applied. Firefly-based clustering is a recent method which proves better for optimal clustering finding. Here, we have combined genetic algorithm with firefly algorithm for optimal clustering. The experimental analysis will be carried out to evaluate the feasibility of the scalable Possibilistic Fuzzy C-Means (PFCM) clustering approach. Finally the analysis was made by two types of datasets the skin and the poker hand data set from UCI machine learning repository. Our proposed method is compared with the performance of the existing genetic and K-Means clustering algorithm. It is analyzed by the accuracy and computation time based on final clusters and varying data sizes. The performance analysis and experimental result showed that our proposed method provide better result. Also the experimental analysis showed that the proposed approach obtained upper head over existing method in terms of accuracy and time. The highest accuracy achieved by the proposed approach is 94%.

- ✓ **A Novel Hybrid Approach in Scalable Parallel Clustering using Firefly Algorithm**

K-Means clustering is a common and simple approach for data clustering but this method has some limitation such as local optimal convergence and initial point sensibility. Intuitively, bio-inspired optimization algorithms should overcome the limitations of K-Means clustering algorithm in finding globally optimum clusters. This research mainly aims to achieve the realization of the limitations of the existing K- Means algorithm and to overcome this limitation we propose parallelization of K-Means using Firefly based clustering method. The proposed methodology introduces a new parallel architecture which can handle large number of clusters. Firefly algorithm achieves this by calculating the initial optimal cluster centroid and then initial K-Means algorithm with optimized centroid to refine them and improves clustering accuracy. The final convergence issue is also addressed and solved to a great extent. Our enhanced algorithm utilizes initial centroid selection strategy which reduces the average number of

iterations used effectively and able to reduce the cost of processing. The experimental results reveal that the new parallel method considerably speedups the computation time. Among the four dataset, Modified algorithm showed better results (higher accuracy) in comparison with the standard Parallel K-Means. It asserts that the algorithm is more efficient and accurate evolved from inherent local and global search of the method. In the Classification efficiency techniques reveals that proposed algorithm got more classification percentage compared with existing algorithms.

The performances of the above algorithms have been studied using standard tools used in industry/academia and they have been experimentally verified. The performance analysis of the proposed methods have been carried out using quantitative and qualitative measures such as execution time, cluster validity indices, clustering accuracy, classification error percentage and classification efficiency. Compared to the conventional as well as existing techniques, the proposed methods give a better performance in terms of quantitative and qualitative parameters.

References

1. Juby Mathew,Vijayakumar R, "Scalable Parallel Clustering Approach for Large Data using Genetic Possibilistic Fuzzy C-Means Algorithm", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC 2014) (Communicated)
2. Juby Mathew and R Vijayakumar, "Scalable Parallel Clustering Approach for Large Data using Parallel K Means and Firefly Algorithm", IEEE International Conference on High Performance and Computing Applications (ICHPCA 2014)(Communicated)
3. Juby Mathew,Vijayakumar R, " Scalable Parallel Clustering using Modified Firefly Algorithm" on IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. I (Dec. 2014), PP 14-24, DOI: 10.9790/0661-16611424.
4. Juby Mathew,Vijayakumar R, "Scalable Parallel Clustering Approach for Large Data using Possibilistic Fuzzy C-Means Algorithm" on International Journal of Computer Applications (IJCA)(0975 – 8887) Volume 103 – No.9, October 2014,PP 24-29. DOI: 10.5120/18103-9195
5. Juby Mathew,Vijayakumar R, "A New Hybrid Scalable Parallel Clustering Approach For Large Data Using FCM And FA" on International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) e-ISSN: 2320-9801,p-ISSN: 2320-9798, Vol.2, Special Issue 5, October 2014,PP 269-276.